Library Resources & Technical Services, 2002, Volume 46, Number 2, Pages 39-49.
ISSN: 0024-2527
www.ala.org
http://www.ala.org/ala/mgrps/divs/alcts/resources/lrts/index.cfm

Characteristics of Resources Represented in the OCLC CORC Database

Tschera Harkness Connell and Chandra Prabha

More and more libraries are providing access to Web resources through OCLC's Cooperative Online Resource Catalog (CORC) and, by extension, OCLC's WorldCat database. The ability to use a database to its maximum potential depends upon understanding what a database contains and the guidelines for its construction. This study examines the characteristics of Web resources in CORC in terms of their subject matter, the source of the content, publication patterns, and the units of information chosen for representation in the database.

The majority of the 414 resources in the sample belonged to the social sciences. Academic libraries and government agencies contributed more than 90% of the records for resources in the sample. Using the Anglo-American Cataloguing Rules, 2d edition (AACR2) definitions for publication patterns that are part of the upcoming 2002 amendments reveals that nearly half of the sample fell into the category of integrating resources. Identifying units of representation of the resources described was more difficult. Existing definitions for Web units in development are not adequate to describe all of the resources in the sample. In addition, there is wide variability in the units of representation chosen for inclusion by the libraries contributing records, resulting in little predictability in what units of information might be found in the database.

One way for libraries to provide access to Web resources is simply to provide a connection to the Internet from public terminals. More and more, however, library staff are providing more than a connection. They are providing enhanced access by organizing and presenting those resources that they consider particularly useful to their users in ways that will help users find them. Some libraries are providing access through the library's Web page, using the library Web page as a portal for resources selected by traditional selection criteria Others are providing access by including records representing Web resources in the online catalog so that users can find items covering the same subject matter, in all formats, from a single database. Many libraries are doing a combination of both.

One aid to librarians wishing to provide access to Web resources through the catalog is the Online Computer Library Center (OCLC) Cooperative Online Resource Catalog service (CORC). For end-users, CORC is a subset of OCLC's Online Union Catalog, WorldCat, which offers descriptions and holdings information for millions of resources in all physical formats. Descriptions are contributed by participating libraries. The CORC portion of the database presents bibliographic records and pathfinders representing electronic resources. The bibliographic records are descriptions of electronic resources; the pathfinders are subject guides of resources on a topic. From a library processing point of view, CORC is a system for creating metadata to describe electronic resources. It also allows the metadata creator to choose from several encoding formats such as MARC, RDF, and HTML meta tags. If a record for a resource is in CORC, CORC works similarly to other OCLC input software in that the person processing the resource can copy catalog and export die record into a local system. However, if there is no record for die resource in CORC, die CORC software creates die basic record. Inputting staff must provide a URL for die resource and choose from die offered metadata formats one to be used for die description. CORC then automatically creates a basic record for the resource, using software to harvest information from the resource itself. Once the basic record is created, staff edit the record and export it to the local

system.

This article reports on a part of ongoing research at OCLC. This part of the project was a joint project of OCLC and die Ohio State University Libraries. The OCLC Web Characterization Project (http//wcp.oclc.org) addresses basic questions about the Web-how big it is, what it contains, how it is evolving. This project examined the characteristics of Web resources that have been identified, evaluated, selected, and described by librarians in the OCLC CORC database. The specific goal of the research reported here was to determine the nature of Web resources described through CORC in terms of their publication patterns and their units of representation. The unit of representation is the level at which die library represents a chosen resource that has a hierarchical relationship to other resources. The publication pattern of a bibliographic resource refers to the completeness or projected completeness of die resource at die time it is released (that is, published). This article also examines die subject matter and source of the resources. The term "source" is used to describe die origin of die Web resource and to describe die library or information agency contributing die descriptive record for die Web resource. Our examination of source determined whether the institution creating the description of the resource was the same institution that had made the Web resource available on the Web. Resources made available and cataloged by the same institution were categorized as internal resources of the contributing institution.

Background

Publication Patterns

In cataloging, a resource that is intended to be complete in a finite number of releases has been considered monographic. Cataloging codes and practice have been less clear in defining nonmonographic resources. Monographic publications are commonly contrasted with serials on the basis that serials continue indefinitely. However, the Anglo-American Cataloguing Rules, 2d edition (AACR2), definition of a serial includes an additional dimension that is unrelated to completeness or time: a serial must be issued in successive parts *(Anglo-American Cataloguing Rules* 1998, 622). Resources that do not meet the added criterion of the serials definition (e.g., loose-leaf publications) are difficult to catalog because they are largely ignored in AACR2. In the environment of tangible formats, these types of publications are proportionately few and catalogers have developed means to work around the lack of guidelines on how to handle them. In the electronic environment, however, the number of resources that continue indefinitely but are not issued in successive parts is great Electronic resources, although they may continue indefinitely, are also often revised continuously. And as they are revised, their form and content may evolve. Discussions about new definitions of publication patterns developed from die recognition that there is no provision in AACR2 to indicate variances from die serials model of successive parts for publications that continue indefinitely.

Hirons et al. refine definitions of publication patterns by dividing all resources into two categories, finite and continuing. Finite resources "are complete or intended to be completed" (Hirons et al. 1999). Finite resources include monographs. "Continuing resources are those that are intended to be continued for an indeterminate period (e.g., serials, updating loose-leaf publications, databases, etc.)" (Hirons et al. 1999). Building on the work of Hirons and others, the Joint Steering Committee (JSC) for die Revision of AACR provisionally approved die definition of continuing resource for addition to AACR2 noting that "[continuing resources include serials and ongoing integrating resources" (Joint Steering Committee 2001). The JSC defines an

integrating resource as a "bibliographic resource that is added to or changed by means of updates that do not remain discrete and are integrated into the whole. Examples of integrating resources include updating loose-leafs and updating Web sites . . . " (Joint Steering Committee 2001). A serial is a "continuing resource issued in a succession of discrete parts, usually bearing numbering, that has no predetermined conclusion. Examples of serials include journals, magazines, electronic journals, continuing directories, annual reports, newspapers, and monographic series" (Joint Steering Committee 2001).

Units of Representation

The importance of indicating the unit of representation when describing the design of bibliographic instruments (e.g., the online catalog), has been well stated by other writers. Wilson calls for makers of bibliographic instruments to state the design specifications of each instrument so that users will be able to get maximum benefit from its use (Wilson [1968] 1978). He notes that "[t]here is a distinction between not finding what we are looking for and finding that what we are looking for is not there; the former is a failure, the latter a negative success" (Wilson, 59). Further, he states that without knowing the "specifications" for the design of the database, it is not possible for the user to make a distinction between the two (Wilson). Bates, in writing about standards for systematic bibliography (which includes the library catalog or database), states similarly that"... the bibliography should not only list materials, but also state information that enables the bibliography to be located *relative to the rest of the graphic universe.* In order to accomplish the latter, we must state precisely what is and is not covered in the bibliography ..." [emphasis in the original] (Bates 1976, 13). While Wilson and Bates differ slightly in their respective lists of required specifications, both consider it essential to define the units that are represented in the database. Bates refers to these units as bibliographic units (Bates 1976, 14), whereas Wilson writes of the unit of representation or "the unit for listing and description" (Wilson [1968] 1978, 61). For the remainder of this article, we will refer to the specification for the unit as the *unit described* or the *unit of representation.*

For tangible resources, the unit described in library catalogs has been determined, in effect, by publishers' packaging and libraries' collection development policies. The issue of the unit of representation has never been well addressed by cataloging codes. What libraries have acquired (book, serial, video, etc.) is what has been described. Individual libraries have had the option to describe groups of books, such as those in a series, instead of each individual book, but the decision is often made on the basis of publisher presentation and/or local collection development policy. If the publisher provides an individual title and numbering for each item in the series, then the library is more likely to describe the individual items. Or, if the library plans to buy all or most of the series, the series may be cataloged as a unit, especially if there are many items in the series and the series is well known. The library may also decide to provide some additional description and access to selected parts, but again that is a local decision.

In the general introduction to AACR2 (1998), the issue of what unit to describe is sidestepped by the following statement: The rules cover the description of, and the provision of access points for, all library materials commonly collected at the present time" *(Anglo-American Cataloguing Rules* 1998, 1). Materials "commonly collected" are the domain, the universe from which materials are selected for inclusion in the database. For these materials, the issue of what unit to represent or describe is addressed in the scope notes of chapters devoted to a particular type of publication. The scope notes set parameters for the type of material covered by the chapter and

in doing so, define the unit to be described under the rules of that particular chapter. For example, chapter 2 presents the rules for describing separately published monographic printed items (i.e., books, pamphlets, and printed sheets). The chief source of information for cataloging these items is the title page. Following the guidelines of this chapter means that units described are whole books, whole pamphlets, and entire printed sheets. Other examples of the units of library materials to be described include whole sound discs and tapes, whole movies and videos, and whole runs of serials.

AACR2 does provide a means for analysis or, "preparing a bibliographic record that describes a part or parts of an item . . ." *(Anglo-American Cataloguing Rules* 1998, 299). However, in practice, analysis is infrequently done. Cataloging a chapter in a book, a single reading from a sound disc, or the music from a motion picture requires a great amount of effort on the part of the cataloging agency. In terms of overall design of online catalogs, AACR2 and common practice for choosing whole units for representation result in a database of resources represented broadly, or stated differently, a database with low granularity in terms of the information units described.

The organizational tradition for archival material also takes a broad approach. Modern archival science is based partially on the assumption that the significance of archival materials "is heavily dependent on the context of their creation, i.e., their *provenance*..." (Hensen 1989, 4). The consequence of this principle is that the cataloging manual *Archives, Personal Papers, and Manuscripts* (APPM) "approaches the problems of archival cataloging principally at the collection level.... [To emphasize] individual components at the expense of the whole collection may tend to obscure the intrinsic importance of the whole" (Hensen 1989, 5). In the scope note for the chapter on description, the APPM provides a list of materials that a collection may contain: correspondence, memoranda, photographs, maps, drawings, pamphlets, broadsides, newspaper clippings, motion picture films, and computer files (Hensen 1989, 9).

One of the difficulties in cataloging new materials is this issue of what to represent. Although the introduction to AACR2 states that the rules can be used "as a basis for cataloging uncommonly collected materials of all kinds and *library materials yet unknown" (Anglo-American Cataloguing Rules* 1998, 1), consensus on the unit of representation for new materials has to evolve (emphasis added). If the new materials are not that different from other materials for which conventions have been established, then consensus maybe quick to form (for example, videocassettes and CD-ROMs have parallels in film and 33 1/3 rpm sound recordings). For Web resources that are digital versions of printed/paper documents, or serials, librarians have tended to choose die same unit of representation as they have for the print counterparts. However, Web resources such as Web sites are not mirrors of tangible resources, and the need for clear definitions has been recognized. The identification of meaningful, distinct Web bibliographic units was a fundamental issue for bibliographic control of Web resources (O'Neill and Lavoie 2000). They also suggested a framework for definitions: "Rather than corresponding to physical objects, meaningful bibliographic units on the Web are found within the structure of Web-accessible information…If their use is complemented by unambiguous definitions, Web sites and Web pages represent useful concepts for identifying bibliographic units on the Web" (O'Neill and Lavoie 2000, 55). They proposed the following definitions for Web page, Web site, and Web collection based on the structure of URLs:

> *Web page:* A distinct information unit composed of one or more HTTP-accessible files, referenced and accessed in its entirety by a single URL (O'Neill and Lavoie 2000, 57).

Web site: A collection of interlinked Web pages residing at the same Web host (59).

Web collection: A portion of a Web site, consisting of multiple Web pages, that represents a distinct resource (59).

O'Neill and Lavoie's definitions are built partially on the work of the World Wide Web Consortium (W3C). Lavoie participated in the Web characterization activities of the W3C that resulted in a 1999 working draft document, *Web Characterization Terminology and Definitions Sheet.* Although no longer an active document of die W3C, this document provides some additional practical definitions for Web resources. The W3C definitions of Web site publisher and Web subsite, in addition to the definitions of a page, a site, and collection, that were refined by O'Neill and Lavoie, have been used for die research reported here. A Web site publisher is a "[p]erson or corporate body that is the primary claimant to the rewards or benefits resulting from usage of the Web site, incurs at least part of the costs necessary to produce and distribute the site, and exercises editorial control over the finished form of the Web site and its content" (Lavoie and Nielsen 1999). A subsite is a "[c]luster of Web pages within a Web site, that is maintained by a different publisher than that of die parent Web site, or host site. The subsite publisher exercises editorial control over the Web pages comprising die subsite, perhaps restrained by some broad guidelines imposed by die host site publisher" (Lavoie and Nielsen 1999).

Method

In preparation for this study, a pilot was conducted using records randomly selected from those entered into the CORC database from October through December 1999. The principal purposes of the pilot were to develop a standard methodology for examining sites and to determine which characteristics of sites would be used as the focus of the second phase of the project reported here. Specific objectives of the pilot were to test the application of existing characterization schemes for describing distinct Web bibliographic units, and to categorize the subject content of those units, the institutional origins of the content of those units, and the institutional sources of records describing those units.

The second phase of the project involved a proportional sample of member-created records, taken over the 12 months of July 1, 1999-June 30, 2000. A sample size of 384 records ([$n = (1.96)^2(.5)^2/(.05)^2$]) was needed for a 95% confidence level. An additional 77 records were drawn for the sample (461 total) so that NetFirst and InterCat records could be eliminated and still meet the needed sample size. A sample of 461 accounts for the possibility that 20% of records would be nonmember records: [(384 + (384)(.20)) = 461].) NetFirst records were eliminated because they are created by OCLC, not member libraries. InterCat records were eliminated because although they are created by OCLC member libraries, they are not created using CORC. After eliminating nonmember records and records for which no usable URL could be determined, die final usable sample was 414 records.

Resources represented by records in the sample and the records themselves were captured on a CD-ROM so that each resource could be examined as it appeared at the time the sample was drawn. In some cases, multiple screen shots of a resource were captured if the Web address accessed a page that served as a collective listing for several different resources, and the bibliographic record described a resource off that page that could not be accessed directly. All resources were then characterized by source, subject matter, publication pattern, and units of

representation.

The characterizations were made by examining each resource. Records were used to assist in die identification of resources only in those cases where a URL was not enough to identify die resource selected by die library. For example, in one case, the URL was to a site that gave a collective tide listing for several agricultural technical reports. Examination of the record revealed that the resource selected by the library was an individual report, not a composite site.

Because we had learned during the pilot project that characterization of the resources in terms of unit described was the most difficult determination for the project, the characterization was performed by several individuals and then discussed in groups. OCLC staff who had been involved in the development of the definitions, as well as the current project team, characterized the resources. In group sessions discrepancies were examined and normalized, if possible, by definition refinement.

Results

Description of the Sample

Contributing Libraries, Internal and External Resources, Subject Matter

As part of the examination of the resources, data were collected on the contributing libraries to determine who was using GORG for cataloging. Academic libraries and government agencies were by far the greatest contributors of records in the sample, contributing a total of 92% of the resource descriptions. Government agencies included national, state, and city governments, governmental departments such as the U.S. Department of Agriculture, governmental regulatory commissions, the military, and law enforcement bodies. Public libraries were not considered government agencies for this study and were counted separately.

Out of 414 records, 67% (278) were contributed by academic libraries. Twenty-five percent (104) were contributed by government agencies, and of those, 23% of die total (94) were contributed by U.S. federal and state agencies. Public libraries contributed only 3% (13) of the records. All other groups (associations/foundations, corporations/business, and networks/consortia) contributed fewer than 10 records each (<3%) (see table 1).

Part of the promise of the Web has been the potential for individuals, groups, and institutions to make available resources that had never been widely available in the past. For that reason the authors were interested in determining to what extent CORC was being used by libraries and other information agencies to describe their own unique resources. In the sample, 21% (88) of die resources were characterized as internal resources and 78% (323) were characterized as external to the institution cataloging. For three resources it was not possible to make a determination.

At first glance, the portion of internal resources (21%) in the sample may seem low; but, given the amount of preparation required to make resources available electronically (e.g., digitization of the resources, database infrastructure creation, metadata assignment, and Web design), it is quite positive that one-fifth of the resources examined in this study were internal or local resources. Said another way, one-fifth of the resources in the sample were "new" resources to the general public. Prior to the Web these resources were only available by traveling to the contributing library or information agency.

Resources in the sample were classed broadly using the Library of Congress classification. The majority of the resources were classed as social sciences (see figure 1). The largest single category, in fact 14% (57/414) of the total sample, was commerce-related. Examples of

commerce-related sites include company and bank Web sites, transportation and commerce regulations, and product catalogs. Other types of social science resources well represented were national, state, and local governmental Web sites. Arts and humanities sites included artifacts of history such as photographs and historic maps, reproductions of paintings, and works of literature. The sciences were represented by sites emphasizing technical issues in agriculture, science, medicine, and military/naval science. Science resources included sites devoted to a particular research project or grant, a particular disease, and even an armed forces technical training curriculum.

Publication Patterns

Using die definitions of Hirons et al. (1999) for finite and continuing resources, 42% (173/414) of the resources in the sample are finite (see table 2). Sixty-nine percent (120/173) of the finite resources mirror traditional monographic resources such as art reproductions, dissertations/theses, books (including exhibit catalogs), and documents. An example of a document is shown in figure 2. The other 31% (53/173) of the finite resources include individual encyclopedia entries, maps, photographs, and archival collections that appear to be complete, for example, the papers of a former university faculty member and department head. Individual photographs such as those from Northwestern University's Curtis collection of historic photographs (see figure 3) make up 215% (44/173) of the finite resources or 11% of the entire sample.

Table 1. Records Contributed to the CORC Database by Library Type (*n*=414)

Contributing Libraries by Type	No. of Records	% of Records
Academic Libraries	278	67.1
Government (U.S.) Libraries	94	22.7
Public Libraries	13	3.1
Government (Non-U.S.) Libraries	10	2.4
Network/Consortia	8	2.0
Association/Foundation Libraries	6	1.4
Corporation/Business Libraries	5	1.2
Total	414	99.9

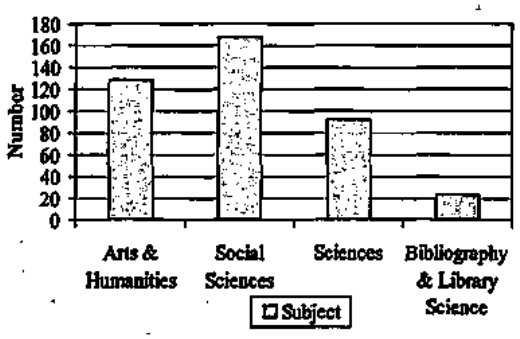

Figure 1. Subject Distribution of CORC Resources (*n*=414)

Continuing resources comprise 58% (241/414) of the sample: 80% (192/241) of these are integrating resources and 20% (49/241) are serials. Overall, serials make up 12% (49/414) of the total sample. Examples of integrating resources include the University of California, Berkeley resources on Iberia (see figure 4) and the Naval Research Laboratory, Chemistry Division home page (see figure 5). These are both integrating resources because, as they are updated, the updates become an integral part of the whole. Unless a snapshot has been archived, there is no way to view the resource as it existed before die update. Figure 6 depicts a serial *(Commission of Preservation & Access Newsletter).*

Table 2. Publication Patterns of Resources (*n*=414)

Publication Patterns	No.	%
Continuing Resources	241	58.2
Integrating Resources	192	46.4
Serials	49	11.8
Finite	173	41.8
Total	414	100.0

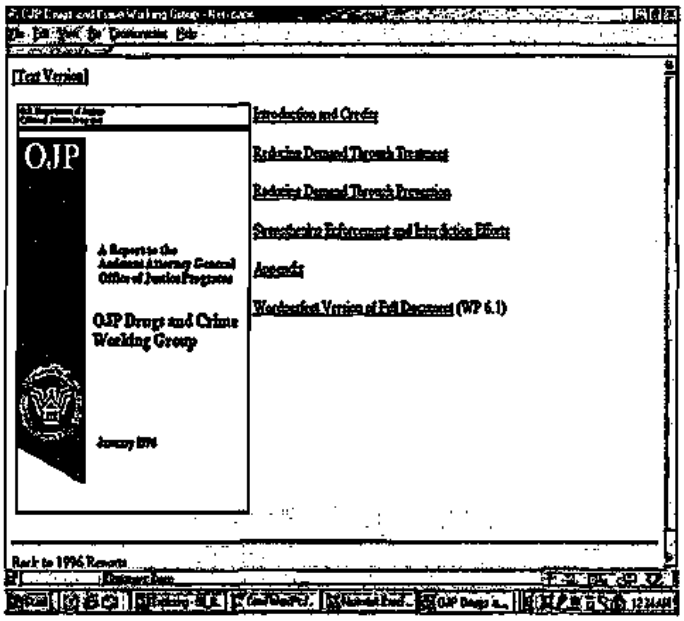

Figure 2. Example of a Monographic Web Resource (Report to the Assistant Attorney General, Office of Justice Programs OJP Drugs and Crimes Working Group, 1996)

To examine units of representation, resources were categorized by two sets of definitions: (1) the traditional physical units of resources in libraries of the twentieth century, and (2) the Web structure units, proposed by W3C (1999) and O'Neill and Lavoie (2000). Using physical unit

definitions involved categorizing the 233 resources (or 56%) that mirrored tangible resources. First, these were categorized by the types of library materials AACR2 presents as *commonly collected.* Within these types, the resources were further broken down in terms of the unit represented, for example, book, chapter, encyclopedia, an entry from an encyclopedia, serial, a single issue of a serial, etc. The 181 resources that were primarily "loose-leaf' in nature were not categorized by the use of tangible resource comparisons.

- *"Resources mirroring tangible resources.* Of the resources of the sample that could be characterized by their tangible counterparts, 63% (147/233) were whole items matching units of materials cataloged in libraries of the twentieth century (see table 3). Examples include reproductions of paintings, whole books, complete databases, dissertations, theses, newspapers, and serials. Thirty-seven percent (86/233) of the materials that mirrored tangible resources would traditionally be considered parts of units and possibly candidates for analysis. In some cases, these resources would not have been described in the catalog, but instead placed in library vertical files. Examples of analytics are 2 entries from an encyclopedia and 15 individual issues from various serials. Twenty of the resources were time-sensitive, similar to brief printed pamphlets or fact sheets-materials traditionally placed in a library vertical file.

- *Categorization of resources using Web-structure definitions.* All but two resources (which were eliminated due to technical difficulties) were characterized by the definitions for Web resources of the W3C and O'Neill and Lavoie (2000). Categorization was accomplished by two groups of individuals working independently. One group categorized all the resources using the W3C definitions for Web sites, subsites, and pages. Individuals in the second group included a category for Web collection. Results are included in table 4. Pages appeared most frequently, totaling more than one-third (35%) of all resources in the sample.

Discussion

Most of the discussion that follows centers on observations and issues relating to the publication patterns and units of representation of the resources. First, however, a few observations about the contributors of the records in the sample. It is not surprising that academic libraries contributed most of the records to CORC in this sample. The data for this study come from records contributed to CORC during FY2000. Prior to July 2000, CORC was still largely experimental and in very active development. In introducing CORC, OCLC used a very different approach than it had historically used for its products and services. CORC was released for public participation early in its design stages in order to encourage evaluation of and collaboration in its development. Many academic librarians and administrators consider it part of their mission to advance the field of library and information science through experimentation and testing of new ideas and are therefore often willing to participate in developmental projects. The sample time frame was the first year that OCLC charged for using the CORC service and promoted it as product rather than prototype. By then many academic libraries had been involved in CORC for some time. Additionally, high participation by academic libraries reflects contributions to OCLC's

WorldCat database as a whole.

Publication Patterns

Data show that more than 40% of the resources in the sample were finite (table 2). Twenty-nine percent (120) fit the traditional images of finite resources, including individual works of art, books, documents, dissertations/theses, law and legislation, and reports. Ten of the resources were actual "electronic books," such as copies of published monographs. By far the largest portion, 25%, of the 173 finite resources in the sample were individual photographs.

Continuing resources made up 58% of the total sample. Most of these (80%) were integrating Web sites; 20% were serials. Overall, serials made up 12% of the entire sample. This figure is actually double the proportion (6%) of serials in WorldCat (OCLC 2001). The disproportionally high numbers of Web sites and individual photographs in the sample may indicate that CORC is being used by libraries primarily for special projects or possibly for experimentation with new types of resources. Web sites and individual photographs are not the types of resources that would easily fit into traditional library work flows. CORC provides a convenient means for trying out new software and work flows and for gaining experience with new types of resources on a project basis.

There were instances when, without knowing the contributing library's intention, categorization of the resource would have been very difficult. For example, figure 7 probably depicts a serial. If the bibliographic resource of interest (to the cataloging agency) is the journal, *Professional Candy Buyer*, the record will represent a serial. If the resource of interest is the Web site that includes the journal as well as other resources for candy buyers, then the record will represent an integrating resource. In this case, the contributing library chose to represent *Professional Candy Buyer* as an integrating resource.

Figure 3. Example of a Monographic Web Resource (Native American Indian Photo, http://hdl.loc.gov/loc.award/lencurt.cp10005)

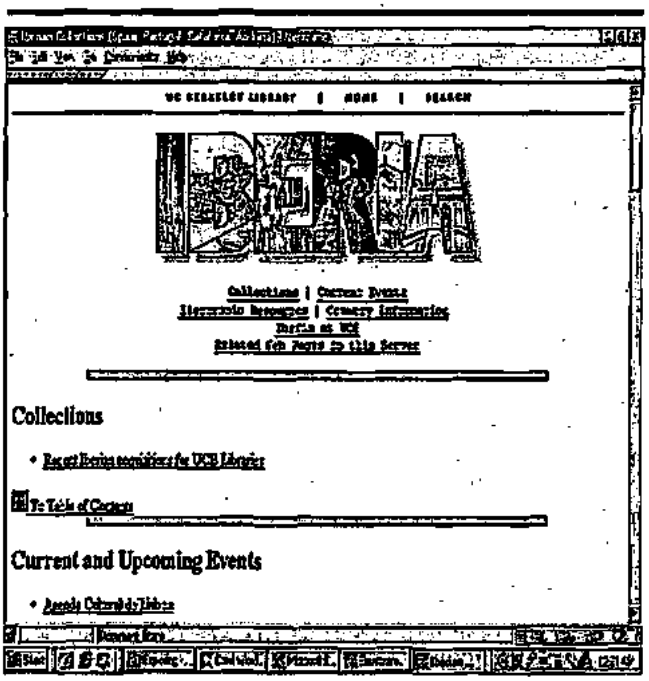

Figure 4. Example of Integrating Web Resource (Iberia, University of California at Berkeley Library, www.lib.berkeley.edu/Collections/Romance/iberia.html)

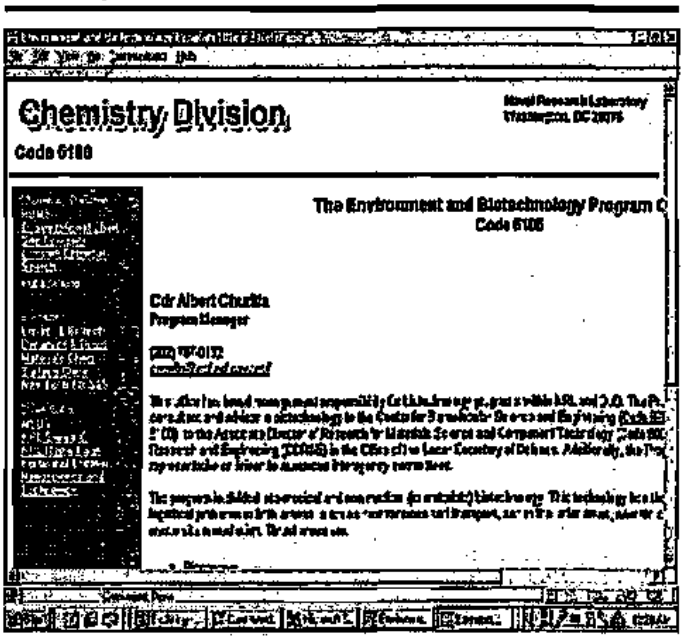

Figure 5. Example of an Integrating Web Resource (Naval Research Laboratory, Chemistry Division. Environment and Biotechnology Office Home Page, www.chemistry.nrl.navy.mil/ 6106/index.html)

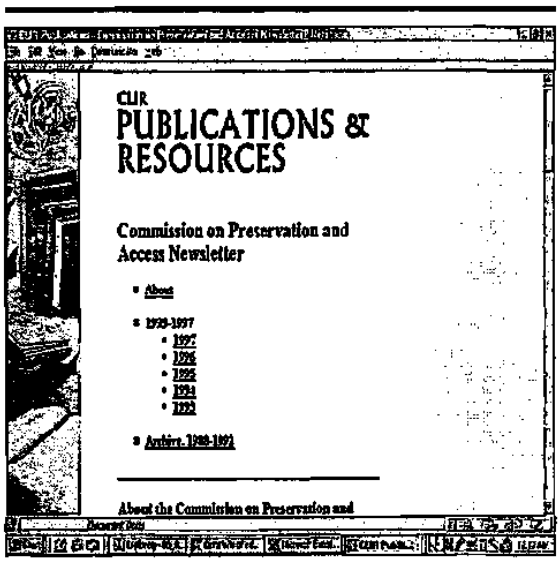

Figure 6. Example of an Integrating Serial Web Resource *(CUR Commission on Preservation and Access International Newsletter,* www.cllr.org/pubs/paln/paln.html)

In this study, statistics were not recorded on the frequency of ambiguous cases such as the *Professional Candy Buyer.* However, ambiguity in how to handle publication patterns is not new. For example, it has long been a local choice to decide how to treat monographic series. Two libraries may choose different solutions for the same series. Any given library will implement different policies for different titles. In some cases the series will be cataloged as a unit, as a serial. In other cases, individual titles in the series will be cataloged separately as finite resources. Over the years libraries have developed guidelines for their local decisions. These guidelines include factors such as whether the library's intent is to purchase the entire series, whether the individual volumes in the series have individual titles, and how the series is treated by indexing services. Also important is how other libraries have handled the series and whether cataloging copy is available. Similar guidelines have yet to evolve for situations such as the one illustrated by the *Professional Candy Buyer.* As guidelines do develop, librarians will be able to provide a level of predictability for users of their catalogs.

Units of Representation in Traditional Terms: Web Resources That Mirror Tangible Resources

Even though more than half (223/414) of the resources in the sample mirror traditional resources, only two-thirds of these (147/223) were represented at unit levels comparable to common practices for handling their tangible counterparts (table 3). The 147 resources that were handled traditionally comprise 36% of the entire sample. Analytics and ephemera make up 21% (86/414) of the sample. Examples of ephemera included an announcement of a town meeting agenda, an advertisement for an upcoming music festival program, and an online "brochure" of free trees available as part of a promotion for Arbor Day. There were numerous instances of photographs that are clearly part of collections of photographs but that were described and represented individually. As discussed earlier, this practice is contrary to archival cataloging principles (Hensen 1989, 5), and while this option is not precluded by AACR2's chapter 8 for

graphic materials, photographs in general purpose libraries have tended to be described as a group. "If the item being described consists of two or more separate physical parts..., treat a container that is the unifying element as the chief source of information ..." *(Anglo-American Cataloguing Rules* 1998, 202). Groups of published photographs or slides are likely to have containers, but following the tradition of archives, even original photographs have been most commonly described as a group (collection) based on general subject matter or provenance. This is especially true when die description of the set is to be integrated into a general topic online catalog. In contrast to common practice, most of the examples of photographs in the sample have been described individually. If a unifying subject has been assigned for purposes of collocation, it has been treated as a series name.

The fact that resources we normally refer to as analytics and ephemeral documents make up 21% (86/414) of the sample means that the resource group we studied is a very different group than the resources represented in a traditional catalog. Said another way, a database for these CORC resources is a very different database in make-up than the online catalog. A fifth of the resources in this CORC sample are small units which contributes to the creation of a database with high granularity; in contrast, the typical online catalog has low granularity. Searching a database with high granularity involves a different set of expectations, search strategies, and vocabulary than does searching a database with low granularity. Patrick Wilson discusses this issue in *Two Kinds of Power*. In the chapter on reliability, he writes of evaluation of bibliographical instruments: "We cannot know how much power is made available to us by a bibliographical instrument unless we know both the plan or Specifications of the work and the quality of workmanship. And each of the separate elements of the Specifications offers a field for the evaluation of performance ..." (Wilson [1968] 1978,127). He considers a number of evaluative questions including, "Have the units to be separately listed been chosen correctly and consistently?" (Wilson, 127). He states that if this question (and the others he has posed) are answered affirmatively, the bibliographic instrument "can then be pronounced reliable or trustworthy.... The overall reliability or trustworthiness of an instrument depends on the exactness and accuracy and consistency with which the rules embodied in the Specifications are applied..." (Wilson, 127).

CORC records become a part of the larger *OCLC* WorldCat database. WorldCat, because of its birth and growth in the last third of the twentieth century, is a traditional database in terms of the units of library resources it represents. Its specifications have been largely governed by the application of AACR2 and other library standards. The results from this study seem to indicate that CORC participants, by contributing records for smaller units, are changing the traditional "specifications" of the WorldCat database. This is not a conscious redefinition of WorldCat; there is very little discussion of what units are to be represented in catalogs. Much of our practice has been formed of habit and tradition. Because WorldCat is so large, the effects of inconsistency in how units of information are represented may not be noticed to any great degree for many years. However, in time the lack of specifications for units in the database could affect users' ability to predict and to find the information they need.

Units of Web Integrating Resources: Web Sites, Subsites, and Pages

At the time this study was designed and characterization of the raw data performed, the working draft "Web Characterization Terminology and Definitions Sheet" (Lavoie and Nielsen 1999) was an active document of the W3C.

Table 3. Level of Representation of Resources That Mirror Tangible Resources, Level of Representation of Monographs and Serials (*n*=233)

	No. of Records	% of All Records in Sample (*n*=414)
Whole units	147	35.5
Art reproductions	5	
Books	10	
Collections	6	
Databases	7	
Dissertations/Theses	5	
Documents	80	
Newspapers	3	
Serials (whole)	30	
Series	1	
Analytics	86	20.8
Documents (ephemeral)	20	
Encyclopedia entries	2	
Maps5		
Photographs (individual)	44	
Serials (single issues)	15	
Total	233	100.0

Table 4. Resource Categories As Indicated by URLs

Web Resource Categories	No. of Records	% of Records
Collections	117	28.3
Sites	85	20.5
Subsites	42	10.1
Pages	146	35.3
Variances or Undetermined	24	5.8
Total	414	100.0

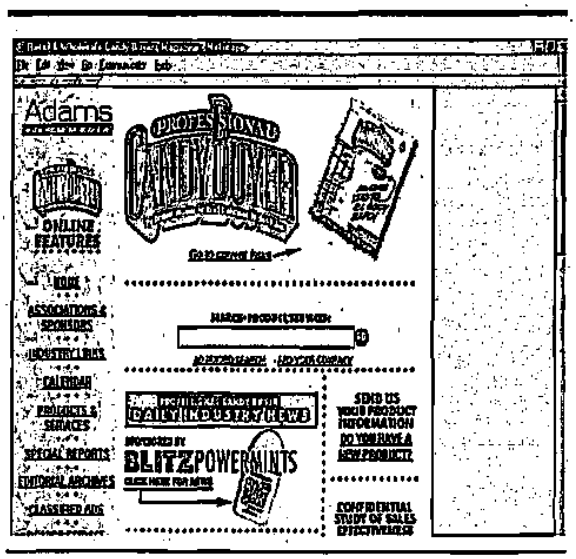

Figure 7. Example of a Serial Integrating Web Resource (Professional Candy Buyer, www.retailmerchahdlslng.net/candy/Default.asp)

Since that time, die document has been dropped as a work item with no evidence on the W3C site that it has been incorporated into another document [www.w3. org/1999/05/WCA-terms/]. A purpose of die working draft was "to bring clarity to the terms often used when talking about the Web" (Lavoie and Nielsen 1999). However, die authors certainly found it difficult to categorize resources according to the definitions.

Categorizing Web sites and Web pages was relatively easy, especially when using O'Neill and Lavoie's (2000) clarifications for site and page. In addition to the base definition that "a site is a collection of interlinked Web pages [or a complete set of Web pages] residing at die same Web host," O'Neill and Lavoie refined the definition by adding that the "access point for the Web site is the home page-the Web page accessed using the base URL of the Web host" (O'Neill and Lavoie 2000,59). Similarly, O'Neill and Lavoie refined the definition of page ("a distinct information unit composed of one or more HTTP-accessible files, referenced and accessed in its entirety by a single URL" (O'Neill and Lavoie 2000,57) by adding two practical considerations:

- A Web page consists of the set of HTTP-accessible files that are viewed simultaneously in a Web browser when the page's URL is accessed (O'Neill and Lavoie 2000,57).
- A Web page located at a given host can be accessed by starting at the host's home page and traversing a sequence of links appearing only in other pages located at the same host (O'Neill and Lavoie 2000,58).

The individuals who categorized the resources in this study had high agreement in their coding of Web sites and Web pages. Two hundred thirty-one resources (55.8%) were assigned to categories of site or page (see table 4). Of these there were only 19 for which categorization differed among those performing the categorization. This translates to 92% (231/250) agreement in categorization of sites and pages.

It was more difficult to categorize those resources that were neither site nor page. The W3C document provided two intermediary categories: subsite and collection, both dependent upon the determination of the role of the publisher of the resource. (For W3C definitions of subsite and Web site publisher, see the "Background" section of this article.) Determination of the role of a corporate entity has never been easy and, in fact, the recognition of this led to the change between the Anglo-American Cataloging Rules (1967) and the Anglo-American Cataloguing Rules, 2d edition (1978) in rules for determining whether a corporate body has principal responsibility for the creation of a resource. AACR2 limits cataloger discretion and only allows assigning principal responsibility for a work to the corporate body under very narrowly defined situations. The authors of this study, in trying to determine the role of die Web site publisher, often referred to AACR2 rules for *construction* of names for corporate bodies. Decisions were made on the basis of whether the corporate body listed on the resource being analyzed as a possible subsite would be considered, according to AACR2 guidelines, as independent of or subordinate to the corporate body listed on the site. O'Neill and Lavoie's (2000) set of definitions were somewhat easier to apply in that they did not include the category *subsite.* They used *collection* as the only intermediary category and defined it independently of the publisher's role: "A portion of a Web site, consisting of multiple Web pages, that represents a distinct resource" (O'Neill and Lavoie 2000,49).

Another difficulty the authors recognized in the assignment of categories is the problem of multiple addresses for a single resource. In the experience of die authors, multiple addresses for a single resource is a situation that arises frequency when describing Web resources, especially

government resources. Because of die design of this research, which captured each resource in time (and place) on a static CD, multiple addresses were not a factor in categorization. However, die issue did arise as researchers re-examined Web sites online to learn more about die resource and its relationship to other sites. The possibility of multiple addresses is a weakness of definitions based on die structure of die Web address.

For the reasons discussed above, die authors were dissatisfied with categorization of resources according to definitions of site, page, and possibly collection and/or subsite. The ease with which definitions can be applied is an important consideration in terms of work flow and productivity, and these definitions were not easy to apply. However, even had they been, the question must be asked as to how meaningful these definitions would be to die general user. Traditional definitions of book, serial, film, etc. have a basis in use as well as a tactile reality. Users carry a book, a serial, or a film. Users do not carry or, we suspect, even routinely search Web resources by the definitions of collection and subsite that were used in this study. Users are likely to think of Web resources in terms of their relationships with other resources. Is the desired site off another? Is it a part of something else? O'Neill and Lavoie's (2000) definitions of site and page do address this navigational aspect of relationship. Unfortunately, only 55.8% of the resources in this study could be categorized as either Web site or Web page.

Conclusions and Suggestions for Further Research

The goals of this research were to determine the nature of Web resources described through CORC in terms of their publication patterns and their units of representation. We also examined the subject matter and source of the resources. The resources in the sample covered the complete range of subjects as represented by the library of Congress classification system. Most resources were contributed by academic libraries, reflecting the contribution patterns of libraries and information agencies to OCLC's larger union catalog, WorldCat. We found it very positive that despite the many practical and financial barriers to digitization, one-fifth of the resources in the sample are unique resources, resources owned only by the information agencies that contributed the record(s) for the resources. Prior to the advent of the Web, these resources would have been unavailable (except through travel) to the general public.

Definitions to aid in the discussion and handling of Web resources are needed so that clear specifications can be developed for databases representing *these* resources. The cataloging community's definitions of finite and continuing resources to describe publication patterns were clear and easy to apply to the resources in the sample. These definitions are important in that they enable the user to predict change in the resource described.

However, definitions for Web units need further development. When AACR2 was written, its rules applied to *materials commonly collected,* which at that time matched the physical packaging of tangible resources. As a result, the unit described for these materials tended to match the content of the package. While this may have been adequate in the environment of tangible, materials, only half (53.6%) of the resources in this sample have a printed/tangible counterpart to aid in their recognition. The concept *of materials commonly collected,* which described the domain of the traditional catalog, is no longer practical as a substitute for clear definitions of units of representation in the catalog.

The definitions for units of representation developed by the W3C and others (Web page, Web subsite, Web collection, and Web site) were also tested, with limited success. The definitions for subsite and collection were difficult to apply and resulted in a great deal of inconsistency

among the results of researchers categorizing resources for this study. The definitions for Web site and Web page were easy to apply, but were applicable for only 55.8% of the resources in the sample. Additional development of meaningful definitions is needed to build databases that provide predictability for the user. The additional definitions should be based on how users use Web resources, how they identify them, how they navigate among them, and how they remember them for future reference. These are issues for further study.

In addition to needing unambiguous definitions for identifying units of representation, there is a need to decide what units will be represented in the database. The data from this study show wide variation in the units of Web resources described by libraries and information agencies. Traditionally, library catalogs represented resources broadly. By contrast, the CORC sample represented a large number of resources that were small units of information (photographs, individual entries from an encyclopedia, Web pages). On the other hand, the sample included large units, such as archive collections, books, serials, and Web sites. This fact raises the issue of the affect of mixing the size of units described in the same database. How do units of representation affect the ability of users to predict potential outcomes of a search and. thus tailor searches for maximum success? This is an important topic for further study.

Works Cited

Anglo-American Cataloguing Rules, 2d ed. 1998 rev. Prepared under the direction of the Joint Steering Committee for Revision of AACR; edited by Michael Gorman and Paul W. Winkler. Chicago: ALA.

Bates, Marcia. 1976. Rigorous systematic bibliography. *RQ* 16, no. 1:7-26.

Hensen, Steven L., comp. 1989. *Archives, personal papers, and manuscripts: A cataloging manual for archival repositories, historical societies, and manuscript libraries.* 2d ed. Chicago: Society of American Archivists.

Hirons, Jean, with the assistance of Regina Reynolds, Judy Kuhagen, and the CONSER AACR Review "Risk Force. April 1999. Revising AACR2 to accommodate seriality: Report to the Joint Steering Committee for the Revision of AACR. Accessed Dec. 16, 2001, www.nlc-bnc.ca/jsc/ser-repO.hbnl.

Joint Steering Committee for the Revision of AACR, 2001. Outcomes of the meeting of the Joint Steering Committee held in Washington, DC, USA, April 2-4, 2001: Revising AACR2 to accomodate seriality. Accessed Dec. 26, 2001, www.nlc-bnc.ca/jsc/0104outhbnl.

Lavoie, Brian, and Henrik Frysryk Nielsen, eds. 1999. *Web characterization terminology* 6- *definitions sheet, W3C working draft 24-May-1999.* Accessed Dec. 7, 2001, www.w3.org/_1999/05/WCA-terms/Ol.

OCLC Online Computer Library Center. 2001. *Annual report 2000/2001.* Dublin, Ohio: OCLC Online Computer Library Center.

O'Neill, Edward T., and Brian L. Lavoie. 2000. Bibliographic control for the Web. *Serials Librarian* 37, no. 2:53-69.

Web characterization terminology and definitions sheet, W3C Working draft May 24, 1999. Brian Lavoie and Henrik Frystyk Nielson, eds. Accessed Dec. 7, 2001, www.w3.oig/1999/0S/_WCA-terms/01.

Wilson, Patrick. [1968] 1978. *Two kinds of power. An essay on bibliographical control* Reprint, Berkeley: Univ. of Calif. Pr.